DOCUMENT RESUME

ED 293 868 TM 011 446

AUTHOR Ervin, Nancy S.
TITLE The Stability of Pretest Statistics Generated from College Samples as a Measure for Predicting Final Form Statistics for Achievement Tests.
SPONS AGENCY College Entrance Examination Board, New York, NY. Admissions Testing Program.
PUB DATE Apr 88
NOTE 44p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *College Entrance Examinations; Difficulty Level; Higher Education; Item Analysis; Predictive Measurement; Pretests Posttests; Regression (Statistics); Sampling; *Statistical Analysis; Test Items
IDENTIFIERS *Admissions Testing Program

ABSTRACT

How accurately deltas (statistics measuring the difficulty of items) established by pre-test populations reflect deltas obtained from final form populations, and the consequent utility of pre-test deltas for constructing final (operational test) forms to meet developed statistical specifications were studied. Data were examined from five subject areas of the Admissions Testing Program (ATP) Achievement Tests (ACH): (1) humanities; (2) mathematics; (3) sciences; (4) social sciences; and (5) foreign languages. The stability of pre-test statistics is evaluated by examining: summary statistics of deltas obtained from pre-test and final form populations; bivariate plots comparing deltas derived from the two populations; the values of the two deltas per item via difference plots; item position shifts; the composition of population samples; and deviation from linearity in the data studied. The research shows that despite the problems inherent in pre-testing on college samples (e.g., the sampling method, motivational factors, relevance of items to course content, method of presentation, item position shifts, and in some cases, inadequate sample sizes), pre-test equated deltas are relatively successful predictors of final form equated deltas. Ten tables and 10 graphs are appended. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED293868

The Stability of Pretest Statistics Generated from College Samples as a Measure for Predicting Final Form Statistics for Achievement Tests [1, 2, 3]

Nancy S. Ervin

Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

NANCY S. ERVIN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

---

[1] A paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, 1988.

[2] The author would like to acknowledge the thorough reviews of this paper by Gary Marco and Linda Cook, the programming and data preparation assistance of Edwin Blew and Susan Bryce and the secretarial assistance of Gaye Allen.

[3] This paper was primarily supported through College Board research funds.

TM 011 446

BEST COPY AVAILABLE

Abstract

The major concern of this investigation is the assessment of how accurately deltas (statistics measuring the difficulty of items) established by pretest populations reflect deltas obtained from final form populations, and the consequent utility of pretest deltas for constructing final (operational test) forms to meet developed statistical specifications. Data is examined from five subject areas of the Admissions Testing Program (ATP) Achievement Tests (ACH): humanities, mathematics, sciences, social sciences and foreign languages. The stability of pretest statistics is evaluated by examination of summary statistics of deltas obtained from pretest and final form populations, study of bivariate plots comparing deltas derived from the two populations, contrasting the values of the two deltas per item via difference plots, investigating item position shifts, evaluating the composition of population samples, and checking for deviation from linearity in the data studied. The research showed that despite the problems inherent in pretesting on college samples (e.g., the sampling method, motivational factors, relevance of items to course content, method of presentation, item position shifts, and in some cases, inadequate sample sizes), pretest equated deltas are relatively successful predictors of final form equated deltas.

The Stability of Pretest Statistics Generated from College Samples as a Measure for Predicting Final Form Statistics for Achievement Tests

As an aid to detecting flaws in items and to obtain reliable item statistics to serve as predictions of how items will perform when used in a final (operational test) form, pretesting represents an integral part of the test development process. The construction of final forms to meet predetermined statistical specifications is largely accomplished by obtaining item statistics generated from pretest samples. In order for pretest item statistics to provide useful estimates for the construction of final forms, these samples must be drawn from a population of test takers that is similar to the population for whom the final form is intended. For many testing programs, this is accomplished by administering pretest items to actual final form test-taking populations by embedding pretest items in final forms or including pretest forms in intact final form variable sections.

Due to a number of considerations (such as available testing time), it is not practical for the College Board Admissions Testing Program (ATP) Achievement Tests (ACH) to pretest items on actual final form populations. For these tests, an alternative practice is carried out in which a particular set of pretest items is administered to college students in a course of the same or comparable materials that the actual final test form is intended to measure. Typically, multiple pretest forms are constructed with each containing a different set of new items to be pretested and each anchored by a common group of items previously administered to final form populations. Each pretest form is administered to college classes, and final forms are then constructed using items from the multiple pretests. The final forms are then

administered to high school students seeking college admission and/or placement in college courses.

A number of important concerns can be raised regarding the appropriateness of pretest data collected from college students:

1) The sampling method: The use of intact classes is a type of cluster sampling. While practical considerations may make such a sampling technique the method of choice, from a statistical perspective it is less defensible than either random or stratified sampling. In addition to requiring broad samples of individuals, cluster sampling imposes the constraint of obtaining broad samples of institutions. (See Henrysson, 1971.)

2) Motivational factors: Unless pretest results directly affect course grades, college students have little motivation to perform to the best of their ability. No scores on pretest data are reported by ATP.

3) Relevance of items to course content: The particular set of pretest items administered to a given college class may not cover material taught in that class.

4) Presentation of pretests: Little information is available regarding the actual manner in which college instructors present ACH pretest materials to their students. Consequently, the extent to which the presentation of a unique pretest varies (in terms of instruction given, adherence to specified time limits, and general explanation of the purpose of the test) across colleges and classrooms within colleges, and the effect of this variability on the item statistics

collected is unknown. See Donlon (1983) for a discussion of this and of the previous two issues.

5) Item position changes: Frequently the position of items within examinations is shifted from pretest to final form. Kingston and Dorans (1984) have found that item statistics may be affected by the position of the items within examinations. There is some concern, therefore, that an item which appears early in a pretest and later in a final form, or vice versa, may exhibit unstable item statistics from pretest to final form.

6) Adequate sample sizes: Eignor & Mazzeo (1985) have noted the difficulty of obtaining a sufficient number of college students to insure the stability of item statistics (particularly results obtained from equating item difficulty indices) as an additional variable affecting the utility of pretest statistics generated on college samples.

The extent to which these factors influence college pretest statistics and their relationship to final form statistics is unclear, although it is likely that they introduce some measure of distortion into indices of item difficulty (deltas) and discrimination (r-biserial correlations).

Test specifications for ACH tests include target means and standard deviations of equated deltas[1] and r-biserial correlations. Because more information is available for item difficulty indices (i.e., targeted means and standard deviations of delta values) than for item discrimination values, and due to time constraints, the study is confined to the investigation of equated delta statistics. Thus, no statement can be made regarding the overall utility of pretesting on non-representative samples.

The major concern of this investigation is the assessment of how accurately pretest deltas generated from college samples reflect final form deltas, and their consequent utility for constructing final (operational test) forms to meet developed statistical specifications.

Stability of Pretest Statistics

Method

Data Source. Pretesting is conducted for all fourteen of the subject examinations in the current ATP Achievement Test battery. Each pretest contains a subset of items previously used within final (operational test) forms, in addition to the new items, in order to establish equated item difficulty (delta) estimates. Pretests are assembled as illustrated below:

Pretest A

Items from: A_1 A_2... A_n X_1 X_2 Y_1 Y_3 Z_1 Z_2

Pretest B

Items from: B_1 B_2... B_n X_1 Y_1 Y_2 Z_1 Z_2

Pretest C

Items from: C_1 C_2... C_n X_1 X_2 Y_1 Y_2 Y_3 Z_1 Z_2

where:

A - C = new, unique items to be pretested as a group within a single pretest form (number varies from form to form); items providing analysis data for this report

X - Z = unique items previously used within operational forms of a test (number varies from form to form); items used for delta equating

Likewise, each new final form is constructed using items from pretests and a subset of items previously used with other final forms.[2] An example of this schematic is:

Final Form D

Items from: A_1 B_6 C_4 B_3 C_2 $\underline{R_3\ R_9\ R_2\ R_6 \ldots R_n}$

where:

A - C = unique items from a pool of pretested items (number varies from form to form); items providing analysis data for this report

R_2 - R_n = unique items previously used within a unique, operational form of a test (number varies from form to form); items used for delta equating

The interdependency of the stability of equated deltas generated by pretest and final form populations is readily observable.

Five subjects from the ACH battery were chosen for investigation in this study: Literature, European History and World Cultures, Mathematics Level II, Physics, and Spanish. These examinations were selected as representative of the defined areas for which tests are assembled for the ATP Achievement Tests (i.e., humanities, mathematics, sciences, social sciences and foreign languages).

With few exceptions, analyses are based on all items for which corresponding pretest and final form equated deltas were available at the time of data collection. The exceptions consisted of:

1) very easy items for which deltas were not calculated (items with a P^+ value greater than .95),

2) very difficult items, for which deltas were not calculated (items with a P^+ value less than .05),

3) items revised between administrations, i.e., items no longer common for the multiple populations,

4) items which were not scored in one or more forms in which they appeared, and

5) items for which P_T (the proportion of examinees who reached a given item) was less than .90 and M_T (the mean delta for the given item based on the responses of only those who reached the item) was not equal to 13.0 (the mean of the delta distribution).[3]

Procedure

For the purpose of this report, stability is defined as the degree of success in using pretest equated deltas to predict final form equated deltas. Summary statistics from equated deltas obtained from pretest and final form populations, bivariate plots comparing equated deltas obtained from pretest and final form populations and difference plots contrasting values of item difficulty determined by pretest and final form populations were examined to assess the stability (after multiple administrations of a form) between pretest and final form equated deltas. Item position shifts were investigated as a possible factor influencing accuracy of prediction. Pretest and final form examinations were divided into quarters and summary statistics were compared for items that had moved to different quarters of the respective tests. Differences between pretest and final form mean deltas by quarter provide a rough estimate of the effect of item position shifts. One possible hypothesis is that items positioned later in the final form than they were in the pretest(s) would appear to be more difficult (i.e., have higher delta values) for the final form population and, conversely, items positioned earlier in the final form would appear to be easier (i.e., have lower delta

values) for that population. It was also hypothesized that the discrepancy between pretest and final form deltas would increase as the discrepancy in the placement of the items (the number of quarters an item was shifted) increased. However, since there is no way of knowing what the delta values would have been if the item position had not been shifted, these effects can not be separated from those due simply to sampling (i.e., the effects of computing deltas based on unrepresentative samples and item position shifts are confounded). Any results obtained from this analysis, therefore, must be interpreted with caution.

Another potentially influential factor (in terms of accuracy of prediction) is differences in populations. These differences were evaluated by examining bivariate plots.

Pretest equated deltas (computed for multiple pretest populations administered different pretest forms) were compared with final form equated deltas (determined for multiple final form populations administered the same final form). The examination was undertaken in order to see if the relationship between the pretest and final form equated deltas is stable. Stability is important for ACH as items from multiple pretest forms are used to build each final form. Once the final form is assembled, it is administered many times to unique final form populations.

To further investigate the stability of equated deltas between pretest and final form populations, difference plots were also produced between pretest equated deltas (as determined within multiple pretest forms) and final form equated deltas (as determined by multiple final form populations administered the identical operational form).

An assumption of the delta equating process is that the relationship between pretest and final form deltas is expected to be linear. Lord (1977), however, notes that the relationship is not strictly linear when the performance of the two groups differs. He also points out that items of different discriminating power can be expected to fall along different curves because they would produce larger differences between the two groups. To investigate the possibility that these two factors might be negatively affecting the stability of pretest to final form deltas, the presence of unexpected deviation from linearity in the data was assessed. This was accomplished by testing a quadratic trend for statistical significance at the .05 level (See Kerlinger & Pedhazur, 1973.) with:

$$F = \frac{(R^2_{y.x,x^2} - R_{y.x})/(k_1 - k_2)}{(1 - R^2_{y.x,x^2})/(N - k_1 - 1)}$$

where: $R^2_{y.x,x^2}$ = the proportion of variance attributable to the quadratic trend,

$R^2_{y.x}$ = the proportion of variance explained by the linear trend, and

$k_1 - k_2$ = the difference in the degrees of freedom associated with the quadratic and linear elements, respectively.

A statistically significant F value would indicate a significant quadratic trend in the relationship between pretest and final form deltas. If such a significant, nontrivial departure from linearity were found, it could help to explain instability of deltas from pretest to final form.

Results and Discussion

Examination of Summary Statistics. For the ATP Achievement Tests, new items are pretested for all fourteen subjects in the battery of tests. The

delta statistics, indicating the difficulty of the items, are used to aid in the construction of new final forms. Thus, stability of these statistics is crucial.

Equated delta summary statistics for new ACH items common from pretest to final forms are presented by subject and form(s) in Tables 1 through 5.

Insert Tables 1 through 5 about here.

The headings of these tables present the number of items common to both pretest and final forms which met the criteria for inclusion in the study (as defined in the data source section). This set of items is not part of the set of common items which was used to establish the equated delta statistics for each item. Final forms of ACH are assembled for administration to multiple populations over a period of time. In addition, final forms are frequently referred to as revised or refurbished forms on the tables. When a few items in a final form need to be modified, the revisions are made and the form is given a K-designation. In recent years, final forms are, after a period of time, refurbished to allow the form to better reflect the evolution of the particular subject being tested. The rows of the tables present the equated delta mean and standard deviation for the set of common items for each testing population as well as the difference between the equated delta means for the items given to the pretest and each final form population. These data reflect the current degree of stability between pretest statistics and final form statistics. In all the tables, a positive difference between final form and pretest mean deltas indicates that equated delta means were higher (items

appeared more difficult) for the final form populations than expected on the basis of pretest equated deltas. Likewise, a negative difference between final form and pretest mean deltas indicates that the items appeared less difficult for final form candidates than expected on the basis of pretest equated deltas. The difference between the pretest and final form equated delta means for the common items ranged from -.7 to -.4 for the Literature forms (Table 1), -.1 to .1 for the Mathematics Level II forms (Table 2), -.8 to -.7 for the Physics forms (Table 3), and .1 to .4 for the Spanish forms (Table 4). There was no difference between the means for the European History and World Cultures forms (Table 5). The overall range of differences between pretest and final form mean equated deltas is -.8 to .4.

For the Literature and Physics Tests, pretest populations tended to overestimate item difficulty from pretest to final form. Questions in the Literature Test are linked to reading passages. The position of items within the test may be affected by the fatigue level of the candidates administered the exam. If the position of items (and their associated reading passages) have been moved forward in the final form from where they occurred in the pretest forms, the movement could be responsible for the change in measures of item difficulty (deltas).

Also, unlike the other subjects in the ACH battery, the construction of a new final form of the Literature Test employs items which have only been administered to pretest populations (i.e., for which equated deltas were obtained for pretest populations using antonym items from retired verbal sections of the College Board Scholastic Aptitude Test (SAT-V) to anchor the measure of item difficulty) as the common item anchor to measure the difficulty of the items in the new test. It is interesting to note, that when

the equating of the refurbished form was carried out, the pattern followed by the other ACH subjects (i.e., using a block of common items which have been administered to final form populations as the anchor to measure the difficulty of the items) was utilized. For the data examined, pretest deltas established by an anchor to SAT-V antonym items and final form deltas established by an anchor to a final form ACH Literature Test were very different.

The Physics data showed the greatest instability of item difficulty values of the five subjects studied as evidenced by the largest discrepancy between pretest and final form mean deltas (i.e., -.8). Physics pretest samples use a background question to select college students who have studied Physics in high school and who are studying Physics in college. For the pretest to multiple final form data studied, all pretest data were determined by candidates just starting their fall semester study of Physics on the college level. Thus, having studied the subject in high school combined with a short period of time elapsing between the completion of high school study and the administration of the pretest form should have fostered a good match for high school students who are nearing the completion of a year of academic study. However, it is possible that the time away from the period of active study combined with the practice of teaching the content of science courses as units had a greater effect for the science area than it did for other subject areas. Cook, Eignor, and Taft (in press) found for another ACH science test, Biology, that the groups of students administered the same items must be at similar time points in their course of study to yield stable equating results. For Physics test items, the time away from active study (and the teaching of the subjects by units) may also have affected the stability of the pretest statistics.

For Spanish forms, pretest populations underestimated item difficulty values as were obtained from final form populations. Perhaps this was because the equated deltas for the final form populations were determined after the native speakers and the candidates studying Spanish less than two semesters had been removed from the group. (As a preliminary to being administered the Spanish Test, candidates are requested to complete a background questionnaire indicating their amount of study of the Spanish Language.) Native speakers were not removed from the pretest samples and all candidates administered the pretests had completed a minimum of two semesters of study of Spanish at the college level. Selection of samples (to use to determine pretest population deltas) based on a background questionnaire for pretest candidates similar to the one utilized by final form populations could result in a better match of pretest and final form samples.

For the Mathematics Level II and the European History and World Cultures forms, pretest to final form deltas were fairly stable.

The European History and World Cultures Test final form equated deltas were determined for items contained in a test measuring all content areas of the European History and World Cultures Test, whereas the pretest deltas were determined for items contained in a subset of the topics covered by the total test. This factor, however, did not seem to impact on the stability of the equated deltas.

Bivariate Plots. Bivariate plots of final form delta values plotted against pretest equated deltas are presented in Figures 1 through 5.

Insert Figures 1 through 5 about here.

For the pretest items administered to multiple final form populations, the final form equated delta values are located on the ordinate and the pretest equated delta values located along the abscissa. For each figure, a different symbol is used to plot deltas obtained from each final form population to show the relationship between pretest and final form equated deltas for multiple final form populations. These plots reflect items for which difficulty indices were first obtained using pretest populations and for which actual values were later determined using data from final form populations. For European History and World Cultures (Figure 1) and Mathematics Level II (Figure 2), the items are fairly equally dispersed on both sides of the 45 degree line indicating that the difficulty of items for these two tests are, in general, fairly stable between pretest and final form.

For the Literature and the Physics Test (Figures 3 and 4), the majority of the items lie below the 45 degree line, indicating that the items appeared to be of greater difficulty for pretest groups than the final form populations found them to be. For the Spanish Tests (Figure 5), the majority of the items lie above the 45 degree line indicating that the items appeared to be of lesser difficulty in the pretest form than the final form populations found them to be.

Difference Plots. Plots of the differences between estimated pretest equated deltas and final form equated deltas are presented in Figures 6 through 10.

Insert Figures 6 through 10 about here.

For the block of pretest items administered to multiple final form populations, the differences between the estimated pretest equated deltas and the final form equated deltas are plotted against a line representing no change in the equated deltas. The measure of the amount of difference between equated deltas is located on the ordinate and the position of the equated delta difference for each pairing of values (for each item) is located along the abscissa. In all of the figures, a unique symbol is used to plot delta differences for each combination of pretest/final form population equated deltas to show the relationship between the measure of difficulty determined for the items by multiple final form populations.

Examination of the difference plots highlights those items with the least amount of stability. In general, across all subjects, once the direction of difference was established (i.e., whether items were easier or harder for the final form populations than they had been for the pretest population), all final form populations moved accordingly. However, for the Literature Test (Figure 6), this motion was intensified for the refurbished form population (indentified on the figure by the symbol *). Items on the Literature Test are linked to reading passages and the measure of difficulty for Literature Test items is expected to be influenced by fatigue of the candidates earlier than for items from tests of other subjects. Thus, change in item position between pretest and final form may lead to greater differences in the measure of item difficulty for Literature items. When items are moved between pretest and final form and then again between final form and refurbished form, the issue is compounded. Repositioning of items by one quarter of the test occurred for 43 percent of the final form items (13 moved forward and 7 moved backward) and 41 percent of the refurbished form items (9 moved forward and 10 moved

backward). Ignoring the data from the refurbished form population, the measure of difficulty for each item remained fairly constant across the three final form populations. Using a criterion of change in equated delta value of greater than 1.0 as reflecting an unstable item, 54 percent of the Literature Test items used unstable pretest values for estimating the actual difficulty of items. Thus, only 46 percent of the measures estimated by pretest populations held for these final form populations.

For the European History and World Cultures Test (Figure 7), 75 percent of the measures of item difficulty held between pretest and final form populations. Looking at the items whose equated deltas were less stable, the values obtained using pretest populations and determined by fall final form candidates (identified on the figure by the symbols X and +) seem slightly more similar than those determined by spring final form candidates (identified on the figure by the * and O symbols).

For the Mathematics Level II Test (Figure 8), 84 percent of the items had an equated difference less than 1.0. Likewise, 73 percent of the items from the Spanish Test (Figure 10) fall into this category.

Over all, 50 percent of the equated deltas changed less than 1.0 for the Physics Test. The direction of change of the equated deltas for the Physics Test (Figure 9), however, was the most consistant across subjects. All items except two, were determined to be easier for all the final form populations than they had been for the pretest population. For these two items, only the spring final form population (identified on the figure by the symbol O) found the items more difficult.

Item Position Shifts. It is hypothesized that item position shifts between pretest and final forms (and sometimes additional item shifts between

final form and refurbished form) may have contributed to the deviation of final form equated delta values from the delta values obtained for pretest populations. Examination was carried out to see how the change in item position was affecting pretest to final form relationships. A rough method of dividing the pretests and the final forms for examination of this variable was employed. For each subject, because the pretest forms and the final forms were not of equal length, each examination was divided into quarters. The degree of item position shift was examined relative to the quarter of the test in which the item resided rather than its position within the test (item number).

Summary statistics of the effects of changing relative item positions from pretest to final form are presented in Tables 6 through 9.

Insert Tables 6 through 9 about here.

Data are provided for items which moved forward one quarter of the test, items which remained in the same quarter of the test, and items which moved backwards one quarter or more of the exam. Because the number of items examined for the Physics Test was insufficient to allow the preparation of meaningful summary statistics, no summary data are provided.

The heading for these tables contains the form definitions. The section labels for the tables contain the number of items moving forward, backward and remaining in the same quarter of each test. Summary statistics are presented along the rows. The correlations between the set of pretest equated deltas and each unique set of final form equated deltas are also reported along the

rows. For data examined across subjects, in general, the hypothesis that items presented earlier in a testing situation are easier for candidates than those presented when a fatigue factor can be present was supported by the data presented. For the Literature Test (Table 6), the decrease in the difference between the means from the multiple administrations ranged from -.7 through 1.1 and for the European History and World Cultures Test (Table 7), the decrease in the difference between the means ranged from -.1 through .2. The items appeared to be easier for the candidates who saw them earlier in the test. If the summary statistics are studied for the effect on item difficulty for items presented later in a testing situation to a second population, the hypothesis that the items would appear to be harder is not borne out across the summary data. However, examination of a block of items tends to underestimate the amount of change made in difficulty that could occur for extreme movement of individual items. An item-by-item investigation across subjects revealed that for the 38 items moved back one quarter or more of the test in final forms from the position they occupied in pretest forms, 63% increased in equated delta value (i.e., the items were harder for the final form candidates). At the same time, for the 37 items presented in earlier quarters of the final forms than they had occurred in pretest forms, 65% decreased in equated delta value (i.e., the items were easier for the final form candidates). Looking at all equated deltas across all subjects, 33% of the delta differences of one or more delta points were for items which had shifted one or more quarters of the test and 100% of the delta differences of two or more delta points were for items which had shifted two or more quarters of the test.

Since the item-by-item examination of change in delta values for items with item placement changes of at least one quarter of the test, in general, showed that the direction of the movement of items could be used to predict the direction of the equated delta change (i.e., equated values decreased for items moved forward in the test and, likewise, values increased for items moved backward in the test), it is suggested that when assembling tests, shifting items for more than one quarter of the test be avoided if at all possible and shifting items for half or three quarters of the test be discontinued.

Population Differences. Using data supplied by the ETS Pretest office, an examination of the pretest populations supplying candidate item responses was undertaken to see if composition of the population sample was a possible factor affecting the computation of pretest equated deltas. Data for the pretest populations which supplied candidate answer responses for the forms examined elsewhere in this research were, unfortunately, not available. Therefore, data for one final form (for each subject represented in the study) was randomly selected. This was a reasonable action because for each ACH pretest form administered, a unique grouping of institutions and their corresponding candidate volumes is present. Thus, the following data came from a random sample of the global composition of pretest populations for the five subjects of interest.

The volumes for the pretest populations used to determine the equated deltas for the pretest items in each of the final forms was good. The Literature, European History and World Cultures, and Physics pretests were above the desired minimal number of candidates for these subjects (200 candidates). For the Mathematics Level II, and Spanish Tests, the candidate

volumes were above the desired minimal number (300 candidates) for the majority of the forms. Two of the pretest populations for the Mathematics Level II Test (N-290 and N-276) and two for the Spanish Test (N-245 and N-242) did not contain the desired minimal volume of candidates.

A better consistency between the number of institutions and the number of candidates administered pretests at each of the participating institutions may have helped to improve the stability of the data between forms. From the information supplied by the ETS Pretest Office, examination of the number of participating institutions across forms ranged from 4 to 14 and the number of answer sheets returned from individual institutions ranged from 3 to 135.

Within subject, the number of participating institutions ranged from 7-13 for the Literature pretests, 4-11 for the European History and World Cultures pretests, 9-11 institutions for the Physics pretests, 7-14 for the Mathematics Level II pretests and 5-13 for the Spanish pretest forms. At the same time, the number of answer sheets returned from the individual institutions for the pretest forms, ranged from: 3-127 for the Literature pretests, 11-120 for the European History and World Cultures pretests, 9-63 for the Physics pretests, 4-135 for the Mathematics Level II pretests and 7-88 answer sheets for the Spanish pretest forms.

Investigation of the geographical areas of the United States represented in the presented data revealed, that across subjects, 36% of the one hundred ninety-nine participating institutions came from midwestern states, 28% came from southern states, 20% came from western states, and 16% came from northeastern states.

Within subject, the midwestern area administered 47% of the Literature pretest forms with the remaining 53% being fairly well divided among the

remaining geographical areas. For the European History and Social Studies pretests, the western states provided 41% of the participating institutions while the midwestern states only supplied 7%. Participating institutions from the combined southern and northeastern states were fairly well divided among the remaining 52% for the European History and World Cultures pretest forms. Examination of the pretest populations administered pretest forms for Mathematics Level II found only 6% of the participating institutions located in northeastern states, 13% located in western states and the balance of 81% in the combined midwestern (38%) and southern states (43%). The location of institutions participating for the Physics Test was fairly equally divided for the southern (35%), northeastern (25%) and western states (30%), but the midwestern states had only a 10% representation. The Spanish pretest forms contained the best balance of representation of geographical location for participating institutions administering ACH pretests (midwestern states 34%, southern and northeastern states each 26%, western states 14%).

In studying the list of institutions administering pretest forms, it was discovered that for one third of the data, the number of answer sheets from one institution overbalanced the breakdown of the number of answer sheets returned from the total number of institutions administering a particular pretest form. For individual pretest forms, there is no attempt to control for representativeness of samples in pretest populations. When the number of answer sheets supplied by one institution overbalance the breakdown of the total number of institutions administering a pretest form, it is possible that delta equating results may be different from those obtained if the number of answer sheets obtained from the participating institutions were more equally represented. It is possible that this was the case for some of the pretest

populations included in the study because an overbalance of individual institutional representation was present for several pretest forms in the examination of the random samples. Six Literature pretest forms supplied items utilized by the final form chosen for study. For one of pretest populations, one institution out of the eleven participating contributed 32% of the answer sheets returned and for another pretest form, one institution out of the eight participating contributed 41%. Of the four pretest forms used for the European History and World Cultures operational form, two institutions out of the six participating contributed 68% of the answer sheets returned for one pretest population, while one institution out of the four participating contributed 34% for a second population. Five pretest forms supplied items for the Mathematics Level II final form. The data showed that for one pretest population, eight institutions participated and two returned 57% of the answer sheets; for a second pretest population, seven schools administered the form and one of them supplied 34% of the data analyzed; for a third pretest form, nine institutions returned answer sheets and one of the institutions represented 37% of the total number returned.

It is expected that the composition of the populations being administered pretest forms could be improved by controlling for consistency between the number of participating institutions for each pretest form and the number of candidates administered the pretest form at each participating institution. At the same time, a careful attempt could be made to exercise balance between geographical areas of the United States in the pretest population administered each form. Improved composition of pretest populations could lead to increased stability of pretest statistics.

Deviation from Linearity. Consideration was given to the possibility that curvilinearity in the relationship between pretest and final form equated deltas may have potentially contributed toward the inaccuracy experienced. For each of the five subjects, F ratios (as described in the procedure section) were computed to test for the significance of the quadratic trend. Because the overall relationship between pretest and multiple final form populations was so similar for each subject, curvilinearity was investigated only for the relationship between the equated deltas computed for the pretest populations and the equated deltas obtained from the original administration of the final form. No significant departure from linearity was found for any of the five subjects. Any suggestions of possible curvilinearity in the bivariate plots appeared to be due simply to scatter effects.

With the knowledge that many categorical variables are affecting the stability of the relationship between the pretest and final form delta values, it is difficult to separate out which causal variables are actually predominant across subjects as well as which are the primary ones for an individual subject. Thus, caution should be employed when evaluating the preceeding findings.

Regression Solution To Increase Stability

Method

Linear regression of final form equated deltas on pretest equated deltas was evaluated as a procedure for adjusting pretest deltas to better predict final form deltas. Despite the fact that final forms are constructed from multiple pretests, the possibility that a single regression line might adequately fit the data was investigated.

Procedure

The model used was the equation for a single regression line and is expressed as:

$$Y = a + bx$$

where: b denotes the common slope parameter for the pretest populations. Fitting the model is equivalent to fitting a single regression line to the combined data.

Results and Discussion

For each subject in the study, the regression model was run for the pretest to final form data which were determined from the original use of the pretest items in a final form. (This set of pretest items is not part of the set of common items which was used to establish the equated delta statistics for each item.) It was anticipated that, when the parameters developed from the regressions would be used to adjust the pretest deltas, the resulting adjusted values would be better predictors for the final form equated deltas than the estimated equated deltas had been. If so, the resulting parameters from the regressions could be routinely applied to pretest equated deltas and thus produce better estimates of item difficulty for the development of new test forms and for meeting predetermined statistical specifications. Summary statistics for the regression runs are presented in Table 10.

Insert Table 10 about here.

Examination of the adjusted values for pretest items, on an item-by-item basis indicated positive results. For the Literature Test items, 63% gave equal or slightly better comparisons with final form delta values. Likewise, 94% of the European History and World Culture items, 84% of the Mathematics Level II items, 50% of the Physics items and 87% of the Spanish items gave better comparisons. Constraints of the study permitted only limited investigation toward the usefulness of developing adjustment parameters for ACH pretest statistics. It is suggested that further research with this topic be undertaken. A larger amount of data should be used to determine the base group to utilize in determining parameters which could be routinely applied to the pretest equated deltas. Cross-validation of the parameters should be performed. It is hypothesized that application of the resulting parameters resulting from the linear regression would likely improve the stability of the relationship between pretest and final form population values of item difficulty.

Conclusions

Despite the problems inherent in pretesting on college samples (e.g., the sampling method, motivational factors, relevance of items to course content, method of presentation, item position shifts, and in some cases, inadequate sample sizes), findings indicate that, for the data examined from the ATP Achievement Tests (ACH), pretest equated deltas are relatively successful predictors of final form equated deltas.

The degree of predictive success was better than expected given the above-mentioned problems. Across the five subjects and multiple administrations, differences between pretest and final form mean equated

deltas (derived from a set of items administered to both populations) ranged from -.8 to .4. The ACH subjects with the greatest differences between mean deltas were Literature and Physics. Excluding these two subjects reduces the range of differences to -.5 to .4.

The factors investigated as possible contributors to instability of deltas from pretest to final form were curvilinearity in the data, item position shifts and population differences. Relationships between pretest and final form equated deltas were found to be essentially linear. On the basis of these findings, it appears that linear regression is an appropriate statistical technique to investigate methods of adjusting pretest equated deltas to better predict final form equated deltas. However, the proportion of variance between the adjusted and unadjusted deltas due to factors other than the adjustment of the statistics may limit the value of routine adjustment of pretest deltas.

As for item position shifts, moving items earlier in final forms one or more quarters of the test than they had occurred in pretests generally resulted in decreasing equated delta values. Likewise, shifting items back one or more quarters of the test often resulted in increases of equated delta values. When data were collapsed across tests, the largest differences between pretest and final form mean deltas were found for items which were shifted two or more quarters later when they appeared in the final forms.

It is important to note that the effects of item position shifts are confounded with those of estimation error in the pretest populations. For this reason, these findings are inconclusive. There does appear to be a trend of higher final form deltas, however, for items whose positions were shifted two or more quarters later in the final form than they occurred in the

pretest. It is, therefore, recommended that this practice be avoided to the extent that it is possible to do so.

Differences between pretest populations were found to contribute to prediction inaccuracy. A better consistency between the number of institutions and the number of candidates administered pretests at each of the participating institutions may have helped to improve the stability of the resulting data. Likewise, balanced representation of all geographical areas of the United States in the pretest populations administered each form may have improved stability.

The practice of using SAT-V antonyms to anchor statistics for the ACH Literature Test may not be the best way to estimate the difficulty of items for this subject.

An additional factor affecting the accuracy of pretest deltas as predictors of final form deltas for ACH is that of final form subgroup membership. For foreign languages, final form deltas are calculated on populations from which native speakers of that language have been excluded. Such is not the case for pretest populations and the results are reflected in differences between pretest and final form deltas. For this exam, pretest deltas tended to underestimate final form deltas. It is strongly suggested that the same sampling constraints be used for both pretest and final form populations for foreign language exams.

Linear regression of final form equated deltas on pretest equated deltas was investigated as a tool to increase the precision of prediction for the forms. Constraints of the study permitted only limited investigation for the usefulness of developing adjustment parameters for ACH pretest statistics. It is suggested that further research with this topic be undertaken utilizing

greater amounts of data and involving cross-validation via the application of the resulting parameters to pretest equated deltas from another form.

No attempt to evaluate the relationship of item content (i.e., effects of social changes on items, classification change over time, multiple classification of items, etc.) was undertaken in this research. Although the stability of the measures of item difficulty between pretest and final form populations was, in general, better than expected for all subjects examined, a few items showed discrepancy of one or more delta points. Further study to research if item content is affecting the more discrepant indices of item difficulty should be undertaken.

\glt
\ne\stabst

References

Bleistein, C. A. & Ervin, N. S. (1987). Assessing pretest statistics generated from college samples as a measure for predicting final form statistics. An empirical study. Princeton, N.J.: Educational Testing Service.

Cody, P. C., & Smith, J. K. (1987). Applied Statistics and the SAS Programming Language (2nd ed.). Elsevier Science Publishing Co., Inc., N.Y., N.Y.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates. Journel of Educational Measurement, in press.

Donlon, T. F. (1983). Pretesting in the Advanced Placement Program: Efforts toward a position paper. Unpublished draft. Princeton, NJ: Educational Testing Service.

Dorans, N.J. (1985). Unadjusting the dropout adjustment for observed deltas computed via IANA10 and IANA34 prior to October 1, 1985. Unpublished memorandum. Princeton, NJ: Educational Testing Service.

Eignor, D. R., & Mazzeo, J. (1985). Pretesting on college samples. Unpublished memorandum. Princeton, NJ: Educational Testing Service.

Hecht, L. W., & Swineford, F. (1981). Item Analysis at Educational Testing Service. Princeton, N.J.: Educational Testing Service.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.), Washington, DC: American Council on Education.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.). Basic Problems in Cross-Cultural Psychology. Amsterdam: Swets & Zeitlinger.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple Regression in Behavioral Research. Holt, Rinehart, & Winston, Inc., N.Y., N.Y.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, 8, 147-154.

Tucker, L. R. (1987). Developments in classical item analysis methods. An empirical study (ETS RR 87-46). Princeton, N.J.: Educational Testing Service.

Footnotes

[1] Equated deltas are item statistics obtained from observed values in a sample of examinees which have been adjusted to estimate the values for a reference population. (See Henrysson, 1971.)

[2] Items common to previously administered final forms are used for score equating purposes.

[3] Prior to October 1, 1985, the mean criterion score of the group that reached the item was used to compute the observed delta. The use of this mean, which often differed from 13.0 when the P^T was somewhat less than 100, frequently resulted in an overcorrection. Beginning October 1, 1985, dropout adjustment for deltas was discontinued. (See Dorans, 1985.)

Table 1

ATP Literature Achievement Test: Comparison of Equated Delta Statistics From Common Items Administered to Pretest and Multiple Final Form Populations

Number of Items = 46	Pretest Forms	Final Form 1	Final Form 2	Final Form 3	Refurbished[a] Form
Mean	12.6	12.2	12.1	12.1	11.9
Difference Between Means	-	-.4	-.5	-.5	-.7
S. D.	1.9	2.0	2.0	2.0	2.0

Note. Number of items in total test was 58.

[a] Final form, which after a period of time, was refurbished to allow the form to better reflect the evolution of the particular subject matter being tested.

Table 2

ATP Mathematics Level II Achievement Test: Comparison of Equated Delta Statistics From Common Items Administered to Pretest and Multiple Final Form Population

Number of Items = 19	Pretest Forms	Final Form 1	Final Form 2	Final Form 3
Mean	15.0	15.1	14.9	15.0
Difference Between Means	-	.1	-.1	0
S. D.	1.8	1.8	1.8	1.8

Note. Number of items in total test was 50.

Table 3

ATP Physics Achievement Test: Comparison of Equated Delta Statistics From Common Items Administered to Pretest and Multiple Final Form Populations

Number of Items = 14	Pretest Forms	Final Form 1	Final Form 2	Final Form 3	Refurbished[a] Form
Mean	13.9	13.1	13.1	13.2	13.2
Difference Between Means	-	-.8	-.8	-.7	-.7
S. D.	1.6	1.8	1.9	1.7	1.7

Note. Number of items in total test was 75.

[a] Final form, which after a period of time, was refurbished to allow the form to better reflect the evolution of the particular subject matter being tested.

Table 4

ATP Spanish Achievement Test: Comparison of Equated Delta Statistics From Common Items Administered to Pretest and Multiple Final Form Populations

Number of Items = 15	Pretest Forms	Final Form 1	Final Form 2	Final Form 3	Refurbished[a] Form
Mean	11.1	11.5	11.3	11.4	11.2
Difference Between Means	-	.4	.2	.3	.1
S. D.	2.4	2.4	2.5	2.5	2.5

Note. Number of items in total test was 85.

[a] Final form, which after a period of time, was refurbished to allow the form to better reflect the evolution of the particular subject matter being tested.

Table 5

ATP European History and World Cultures Achievement Test: Comparison of Equated Delta Statistics From Common Items Administered to Pretest and Multiple Final Form Populations

Number of Items = 36	Pretest Forms	Final Form 1	Final Form 2	Final Form 3	Revised[a] Form
Mean	11.7	11.7	11.7	11.7	11.7
Difference Between Means	-	0	0	0	0
S. D.	2.0	1.8	1.9	1.8	1.8

Note. Number of items in total test was 100.

[a] Final form with minor alteration to items (e.g., because of printing errors in test copy).

Table 6

ATP Literature Test: Summary of the Effects of Changing Relative Item Position From Pretest to Final Form

	Pretest Forms	Final Form 1	Final Form 2	Final Form 3	Refurbished[a,b] Form
13 (19) Items Moved Forward One Quarter or More of the Test					
Mean	12.7 (13.1)	12.0	12.0	11.8	12.0
Difference Between Means	-	-.7	-.7	-.9	-1.1
S. D.	2.2 (2.4)	2.0	2.0	2.0	1.3
Correlation	-	.936	.945	.942	.156
26 (27) Items Remained in the Same Quarter of the Test					
Mean	12.6 (12.4)	12.3	12.3	12.3	11.8
Difference Between Means	-	-.3	-.3	-.3	-.6
S. D.	1.9 (1.9)	2.1	2.2	2.2	2.2
Correlation	-	.858	.843	.836	.711
7 (10) Items Moved Backwards One Quarter or More of the Test					
Mean	12.2 (12.7)	12.0	12.0	12.0	12.1
Difference Between Means	-	-.2	-.2	-.2	-.6
S. D.	1.7 (1.7)	1.8	1.9	1.8	2.2
Correlation	-	.878	.856	.868	.469

Note. Number of items in analysis was 46; number of items in total test was 58.

[a] Final form, which after a period of time, was refurbished to allow the form to better reflect the evolution of the particular subject matter being tested.

[b] Pretest statistics for the refurbished form are enclosed in parenthesis.

Table 7

ATP European History and World Cultures Test: Summary of the Effects of Changing Relative Item Position From Pretest to Final Form

	Pretest Forms	Final Form 1	Final Form 2	Final Form 3	Revised[a] Form
17 Items Moved Forward One Quarter or More of the Test					
Mean	12.3	12.1	12.2	12.2	12.1
Difference Between Means	-	-.2	-.1	-.1	-.2
S. D.	1.6	1.1	1.3	1.2	1.4
Correlation	-	.703	.731	.738	.832
8 Items Remained in the Same Quarter of the Test					
Mean	12.8	12.3	12.3	12.3	12.3
Difference Between Means	-	-.5	-.5	-.5	-.5
S. D.	1.4	2.3	2.3	2.0	2.0
Correlation	-	.884	.904	.907	.902
11 Items Moved Backwards One Quarter or More of the Test					
Mean	10.0	10.4	10.6	10.6	10.7
Difference Between Means	-	.4	.6	.6	.7
S. D.	1.8	2.0	2.0	2.0	1.8
Correlation	-	.748	.737	.737	.788

Note. Number of items in analysis was 36; number of items in total test was 100.

[a] Final form with minor alteration to items (e.g., because of printing errors in test copy).

[b] Test statistics for the refurbished form are enclosed in parenthesis.

Table 8

ATP Mathematics Level II Test: Summary of the Effects of Changing Relative Item Position From Pretest to Final Form

	Pretest Forms	Final Form 1	Final Form 2	Final Form 3
3 Items Moved Forward One Quarter or More of the Test				
Means				
Difference Between Means		Not a sufficient number of items to provide meaningful statistics.		
S. D.				
Correlation				
8 Items Remained in the Same Quarter of the Test				
Mean	14.4	14.2	14.0	14.1
Difference Between Means	-	-.2	-.4	-.3
S. D.	2.1	1.8	1.7	2.0
Correlation	-	.954	.955	.943
8 Items Moved Backwards One Quarter or More of the Test				
Mean	15.6	15.5	15.3	15.6
Difference Between Means	-	-.1	-.3	0
S. D.	1.5	1.8	1.7	1.7
Correlation	-	.932	.865	.932

Note. Number of items in analysis was 19; number of items in total test was 50.

[a] Insufficient number of items for reporting statistics.

Table 9

ATP Spanish Test: Summary of the Effects of Changing Relative Item Position From Pretest to Final Form

	Pretest Forms	Final Form 1	Final Form 2	Final Form 3	Refurbished[a,b] Form
No (0) Items Moved Forward One Quarter or More of the Test					
Mean					
Difference Between Means		Not a sufficient number of items to provide meaningful statistics.			
S. D.					
Correlation					
10 (9) Items Remained in the Same Quarter of the Test					
Mean	10.8 (10.6)	11.1	11.0	11.2	10.7
Difference Between Means	-	.3	.2	.3	.1
S. D.	2.2 (2.3)	2.2	2.3	2.3	2.5
Correlation	-	.940	.929	.935	.930
5 (6) Items Moved Backwards One Quarter or More of the Test					
Mean	11.9 (12.0)	12.2	11.9	11.7	12.1
Difference Between Means	-	.3	0	-.2	.1
S. D.	2.9 (2.6)	2.9	3.2	3.1	2.8
Correlation	-	.978	.975	.969	.972

Note. Number of items in analysis was 15; number of items in total test was 85.

[a] Final form, which after a period of time, was refurbished to allow the form to better reflect the evolution of the particular subject matter being tested.

[b] Test statistics for the refurbished form are enclosed in parenthesis.

Table 10

ATP Achievement Tests: Summary Information from Linear Regression Analysis[a]

Subject	Number of Items Adjusted	Number of Items in Total Test	$r^{b}_{y,x}$	a	b
Mathematics Level II	19	50	.91	1.7541	.8747
Literature	46	58	.88	2.2919	.8483
European History and World Cultures	36	100	.79	1.7325	.8558
Physics	14	75	.87	4.0864	.7513
Spanish	15	85	.96	.9647	.0784

[a] Adjustment equation = a(equated delta from pretest population) + b.

[b] Significant at .05 level.

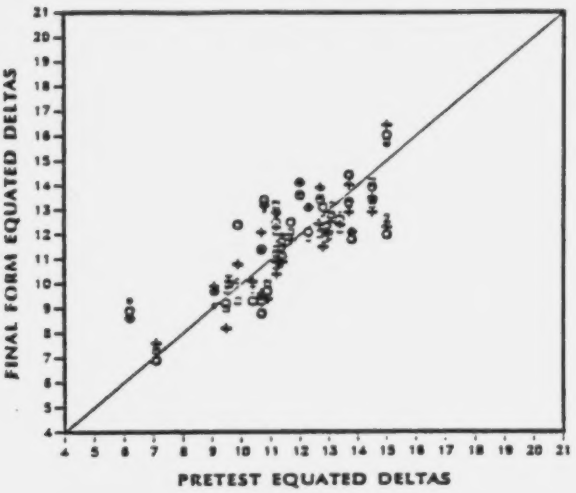

Figure 1: ATP European History and World Cultures Achievement Test (Pretest Forms to Final/Revised Form)

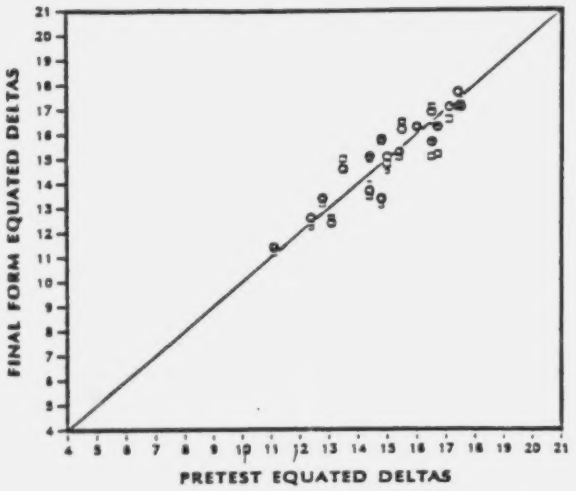

Figure 2: ATP Mathematics Level II Achievement Test (Pretest Forms to Final Form)

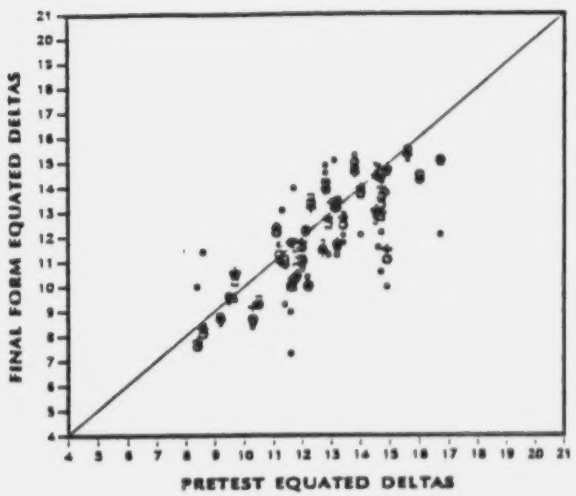

Figure 3: ATP Literature Achievement Test (Pretest Forms to Final/Refurbished Form)

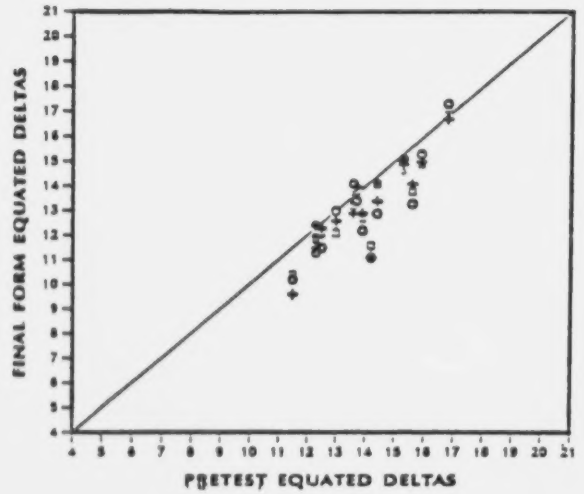

Figure 4: ATP Physics Achievement Test (Pretest Forms to Final/Refurbished Form)

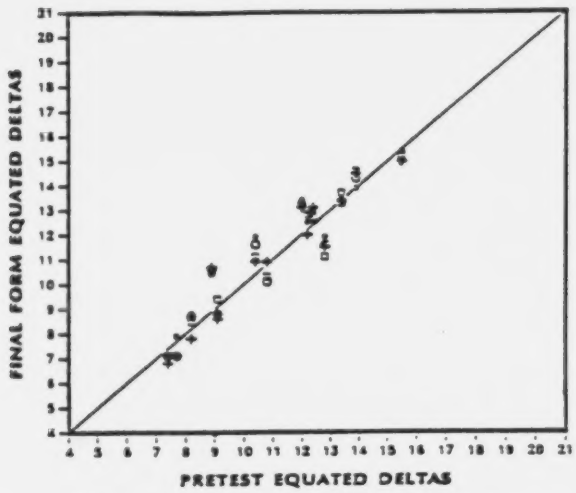

Figure 5: ATP Spanish Achievement Test (Pretest Forms to Final/Revised/Refrubished Form)

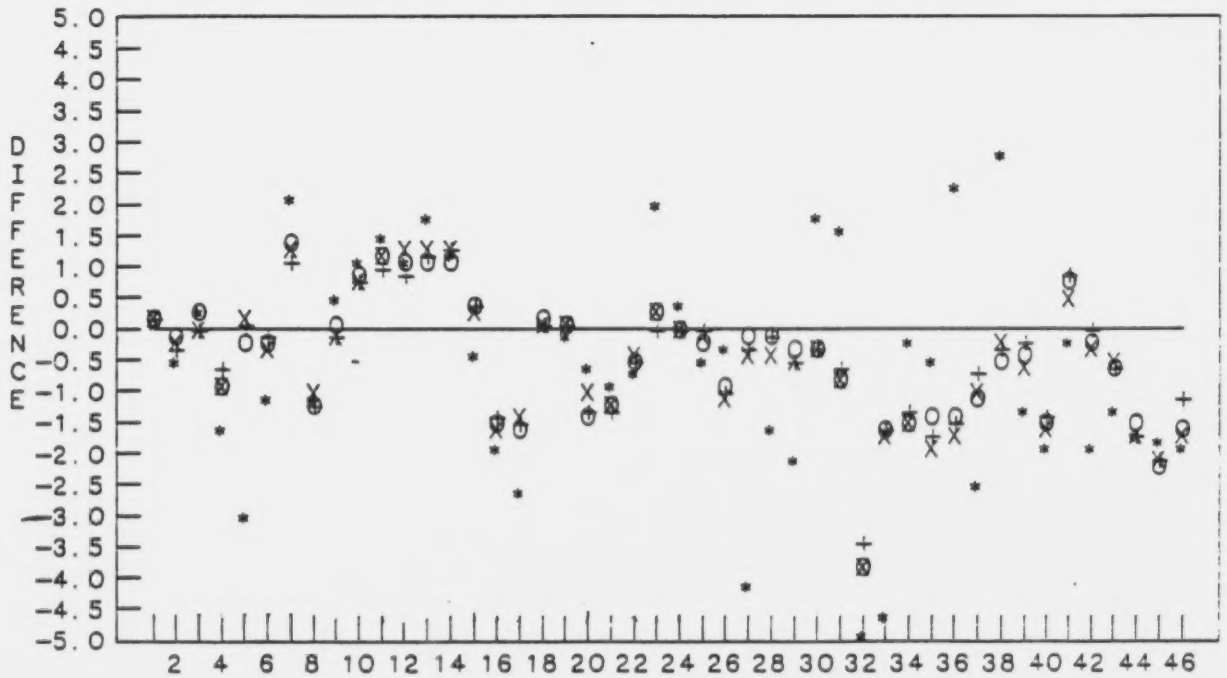

Figure 6: ATP Literature Achievement Test (Pretest to Final/ Refurbished Forms) Equated Deltas

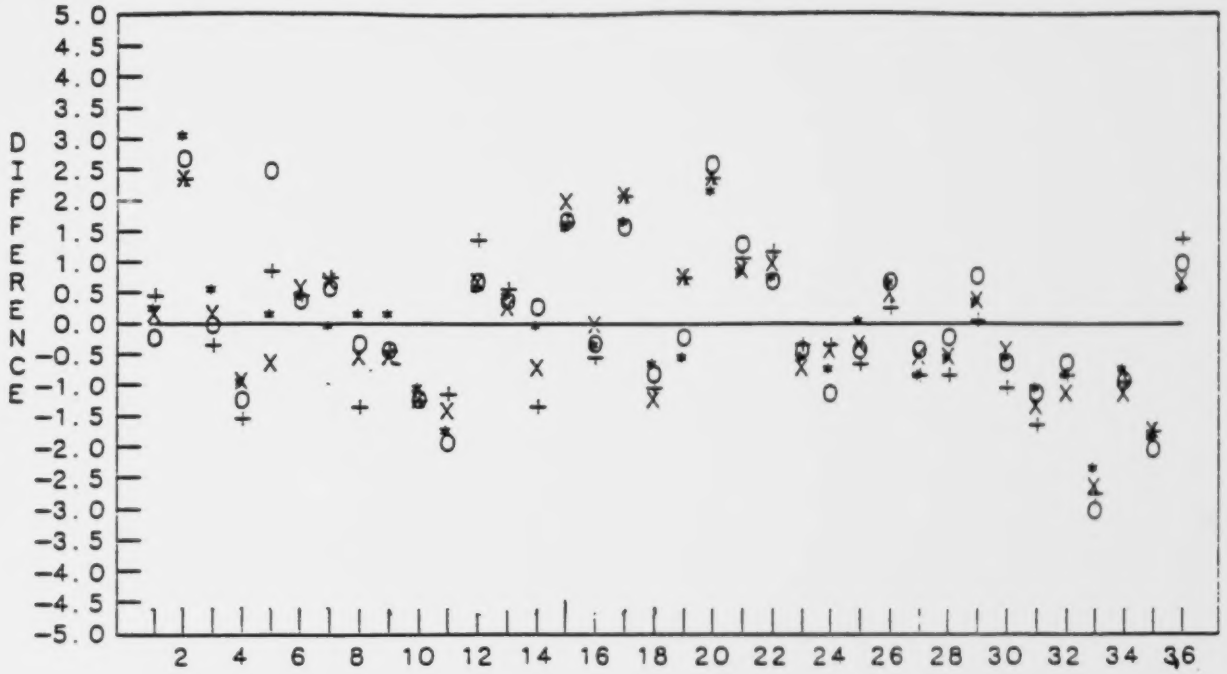

Figure 7: ATP European History and World Cultures Achievement Test (Pretest to Final Forms) Equated Deltas

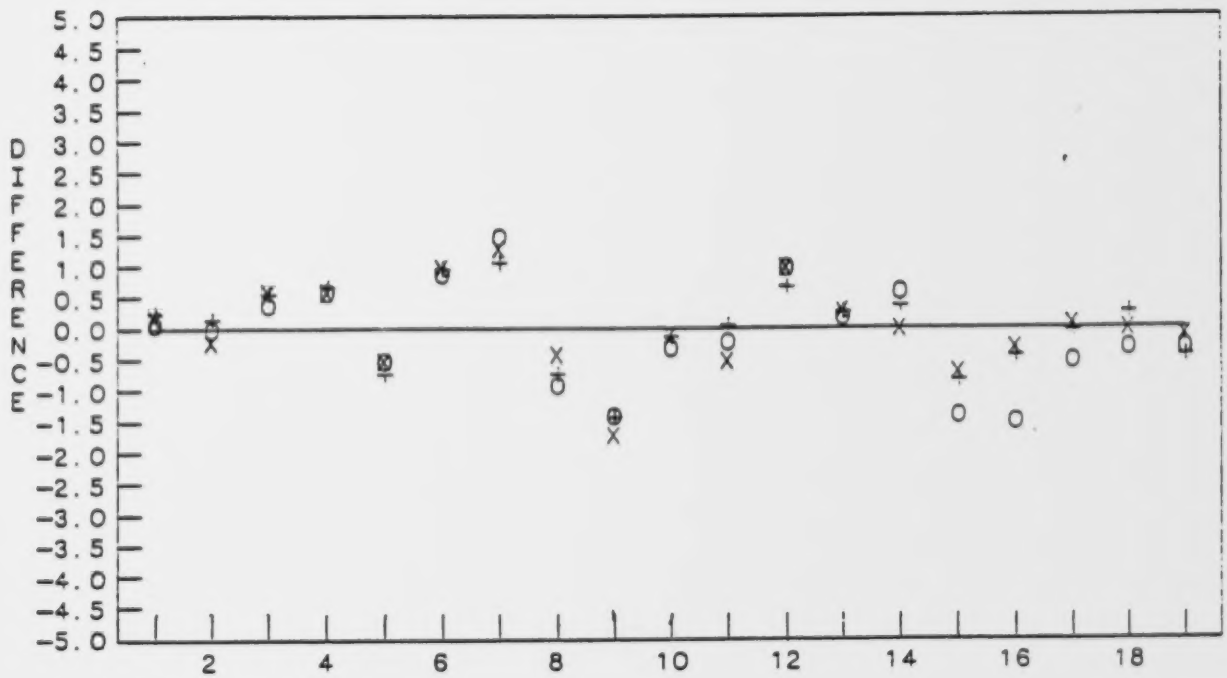

Figure 8: ATP Mathematics Level II Achievement Test (Pretest to Final Forms) Equated Deltas

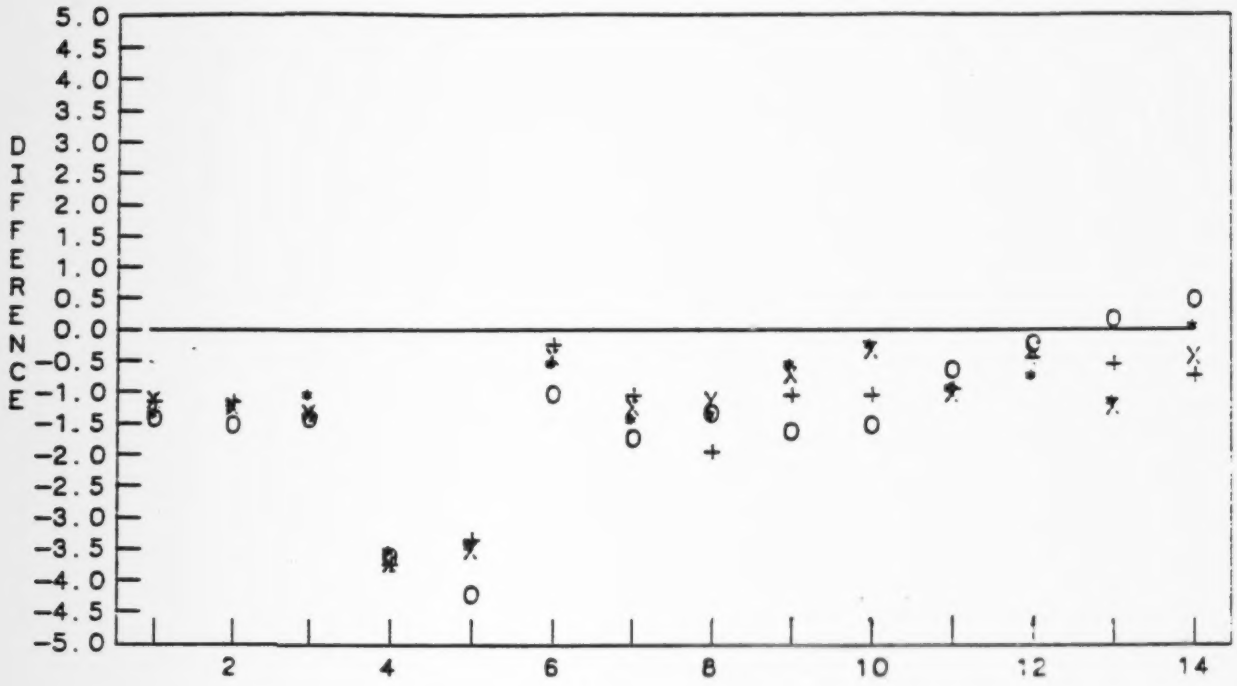

Figure 9: ATP Physics Achievement Test (Pretest to Final Forms) Equated Deltas

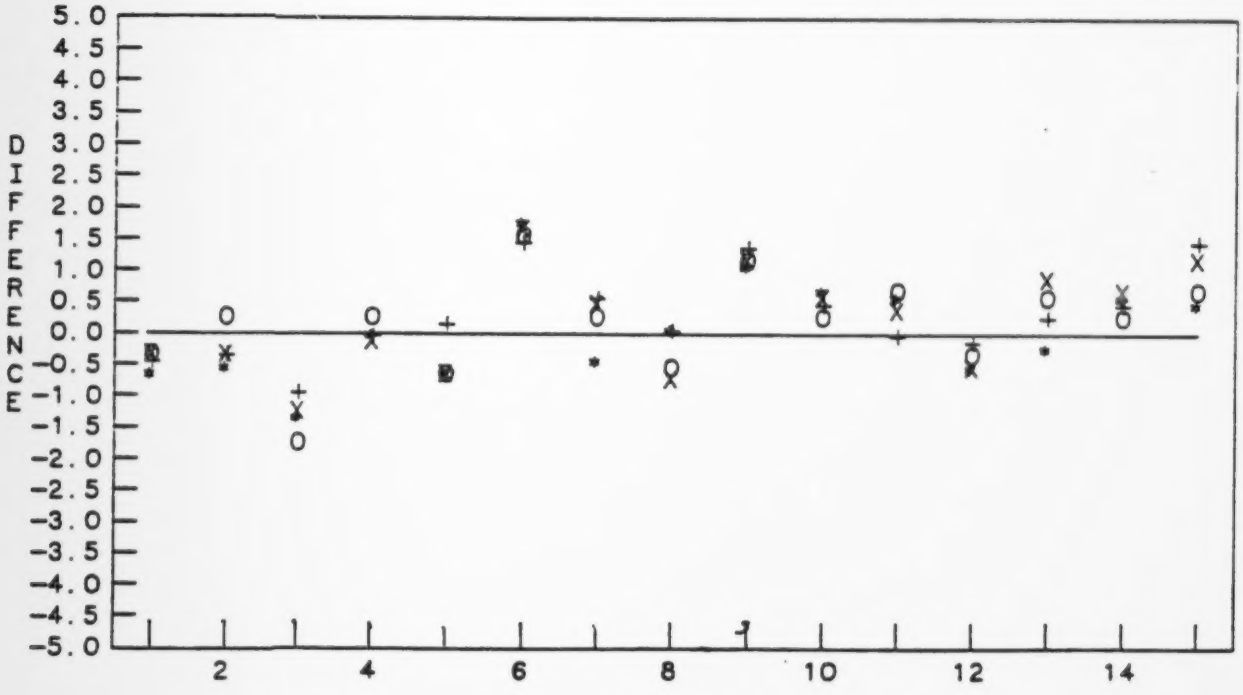

Figure 10: ATP Spanish Achievement Test (Pretest to Final/ Forms) Equated Deltas